(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
5 October 2017 (05.10.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/167628 A1

(51) International Patent Classification:
G06F 19/00 (2011.01) *G06F 17/30* (2006.01)

(21) International Application Number:
PCT/EP2017/056898

(22) International Filing Date:
23 March 2017 (23.03.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
16162540.5 29 March 2016 (29.03.2016) EP

(71) Applicant: **KONINKLIJKE PHILIPS N.V.** [NL/NL]; High Tech Campus 5, 5656 AE Eindhoven (NL).

(72) Inventors: **HULSEN, Tim**; High Tech Campus 5, 5656 AE Eindhoven (NL). **VAN DER LINDEN, Wilhelmus, Petrus, Maria**; High Tech Campus 5, 5656 AE Eindhoven (NL). **PLETEA, Daniel**; High Tech Campus 5, 5656 AE Eindhoven (NL). **OBBINK, Jan, Hendrik**; High Tech Campus 5, 5656 AE Eindhoven (NL). **QUIST, Marcel, Johannes**; High Tech Campus 5, 5656 AE Eindhoven (NL).

(74) Agents: **ZHU, Di** et al.; Philips International B.V. – Intellectual Property & Standards High Tech Campus 5, 5656 AE Eindhoven (NL).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) Title: DATA MODEL MAPPING

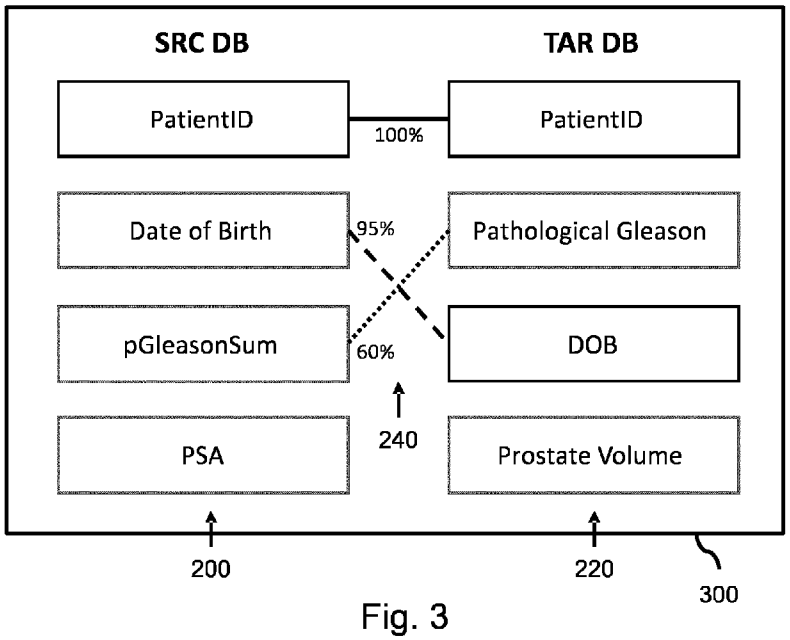

Fig. 3

(57) Abstract: A system and method are provided for generating a mapping function for use in loading data from a source database, which is structured in accordance with a source data model, into a target database, which is structured in accordance with a target data model. The mapping function is automatically generated, e.g., on the basis of descriptions of both data models and/or the data items comprised in the both databases, with the mapping comprising components for mapping respective data elements of the source data model to the target model. To improve the quality of the mapping function, confidence values of the mapping are determined and visualized together with the mapping. User feedback is obtained indicative of a correction of at least one component of the mapping function. In response, the mapping function is adjusted on the basis of the user feedback data. Advantageously, the mapping function may be generated in a 'hybrid' manner, which is less labor intensive and less error prone than having to generate the mapping function by hand.

WO 2017/167628 A1

Data model mapping

FIELD OF THE INVENTION

The invention relates to a system and a method for mapping data models. The invention further relates to a use of the system to extract, transform and load of the data items from a source database into a target database, and to a computer readable medium comprising instructions for causing a processor system to perform the method.

BACKGROUND OF THE INVENTION

There may be a desire to load data items from one database into another database, i.e., from a source database into a target database. For example, it may be desirable to load clinical data from a number of local databases, each belonging to, e.g., a different hospital or other clinical site, into a global database. A specific example may be the Movember GAP3 Global Prostate Cancer Active Surveillance Initiative, where it is said that by integrating clinical, imaging and biomarker data in a global central database, the world's leading research and clinical groups focusing on prostate cancer may be united, so as to allow development of a new therapeutic guideline for men diagnosed with low risk prostate cancer.

However, there often is a mismatch between the data models defining the source database and the data model defining the target database. Namely, the data items in the database may be structured differently, in that the data elements of each database may differ. Here, the term "data element" refers to the logical definition of a field in the database, the term "data item" refers to the actual data stored in the field, and the term "data model" refers to a logical definition of the data elements together structuring the database.

For example, data elements may be differently labeled in both data models. A specific example may be that a data element labeled "*Patient name*" in the source data model may be labeled "*First name, Last name*" in the target data model. It may also occur that the databases comprise similar but not identical data elements. A specific example is that that a data element labeled "*Patient name*" in the source data model may be represented by two data elements in the target data model, e.g., by the data elements "*First name*" and "*Last name*". Yet another example is that a data element labeled "*Weight*" in both databases may represent data items expressed in different units in each of the databases, e.g., in kilograms in

the source database and in grams in the target database. Yet another example is that the data elements in the source data model may be named incorrectly (for example due to a column header shift), have a coded name (e.g. 'source_data_001'), or not named at all.

To correctly load data items from the source database into the target database, it thus may have to be determined first how to map the data elements from the source database to those of the target database. It is known to create such mapping functions by hand. Disadvantageously, creating the mapping function is labor intensive and error prone.

SUMMARY OF THE INVENTION

It would be advantageous to obtain a system and method for mapping data models which is less labor-intensive and/or error prone yet provides a mapping function which is of similar or higher quality as a manually generated mapping function.

A first aspect of the invention provides a system for mapping data models, comprising:

an input interface configured to access:

- a source database structured according to a source data model,
- a target database structured according to a target data model, and

a processor configured to generate a mapping function mapping the source data model to the target data model to enable loading data items from the source database into the target database on the basis of the mapping function, wherein the mapping function comprises components for mapping respective data elements of the source data model to the target model,

wherein the processor is configured for determining confidence values of the mapping represented by respective components of the mapping function as a result of the mapping function being automatically generated; and

a user interface subsystem comprising:

- a display processor configured to generate display data comprising a visualization of the mapping function and the confidence values, and
- a user input interface configured to obtain user feedback data indicative of a correction of at least one component of the mapping function;

and wherein:

the processor is configured to adjust the mapping function on the basis of the user feedback data, and

wherein the processor is configured to generate the mapping function using both data-driven mapping and semantic mapping.

A further aspect of the invention provides a use of the system to extract, transform and load of the data items from the source database into the target database.

A further aspect of the invention provides a method of mapping data models, comprising:

accessing:

- a source database structured according to a source data model,
- a target database structured according to a target data model, and

generating a mapping function using both data-driven mapping and semantic mapping for mapping the source data model to the target data model to enable loading data items from the source database into the target database on the basis of the mapping function, wherein the mapping function comprises components for mapping respective data elements of the source data model to the target model,

determining confidence values of the mapping represented by respective components of the mapping function as a result of the mapping function being automatically generated;

generating display data comprising a visualization of the mapping function and the confidence values;

obtaining user feedback data indicative of a correction of at least one component of the mapping function; and

adjusting the mapping function on the basis of the user feedback data.

A further aspect the invention provides a computer readable medium comprising transitory or non-transitory data representing instructions arranged to cause a processor system to perform the method.

The above measures involve accessing two databases which differ, not only by their contents, e.g., by the data items stored therein, but also by their structure, e.g., by the data elements of each database. As such, there exist differences between the data models modeling the databases. A mapping function is automatically generated. The mapping function comprises components which map individual data elements from the source database to the target database. For example, the mapping function may map a data element labeled “Patient name” from the source database to the data element “Name of Patient” of the target database. The mapping function may be generated as structured data.

However, an entirely automatically generated mapping function is typically of lesser quality than a manually generated mapping function.

To increase the quality of the mapping function, confidence values are determined which represent an estimated match quality of the mapping between data elements by a component of the mapping function, and may thus also be referred to as 'match quality values'. For example, the confidence value may be based on semantic similarity between the data elements, with a mapping of "Patient name" to "Name of Patient" yielding a high confidence value, and a mapping of "Patient name" to "Patient ID number" yielding a low confidence value. Another example is that the confidence value may be determined based on a comparison of data items associated with the data elements. It is noted that such confidence values may already be available as a result of the mapping function being automatically generated, namely by being an internal parameter which is used to select the 'best' mapping for a particular data element, e.g., having the highest match quality. However, the confidence value may also be determined separately, e.g., by comparison of the labels of the data elements, by comparison of the data items associated with the data elements, etc.

The confidence values are displayed to a user together with the components of the mapping function. For example, a line between the visual representations of a source data element and a target data element may represent a component of the mapping function, with a line propriety (e.g., color), additional text or other visual property representing the confidence value. The user is enabled to provide feedback on the mapping function. For example, the user may reject a component of the mapping function, or correct the mapping component. Based on the user feedback, the mapping function is then automatically adjusted.

Advantageously, generating a mapping function by using both semantic mapping and data-driven mapping together will result in a decrease of false positives and false negatives, and is more reliable than using only one of these two different mappings.

Optionally, the processor is configured to generate the mapping function on the basis of a source data model description describing the source data model and a target data model description describing the target data model. The data model descriptions may facilitate semantic mapping between data elements of both databases, in that they may contain, or be indicative of, identifiers of the data elements, e.g., names, labels, identification codes, etc. Such description may be accessed as structured data, e.g., as XML, and is also referred to simply as a "codebook" defining a structure of the database.

Optionally, the processor is configured to generate the mapping function using data-driven mapping and semantic mapping. Here, data-driven mapping refers to a mapping which is generated based on the data items associated with a data element.

Such the data-driven mapping may be based on at least one of:

- a value of a data item of a source data element,
- a relationship between the value of the data item of the source data element and data items of one or more other source data elements,
- a range of values of data items of the source data element, and
- a distribution of values of data items of the source data element,

being indicative of a target data element.

Optionally, the processor is configured to, when generating the mapping function, use data-driven mapping for mapping a source data element to a target data element when the semantic mapping does not meet a mapping criterion. It has been found that a combination of data-driven mapping and semantic mapping is particularly advantageous. Namely, in case the semantic mapping yields a high confidence value, such semantic mapping may normally be relied on. However, if the semantic mapping yields a low confidence value, data-driven mapping may additionally be employed. In addition, data-driven mapping may be used to verify the correctness of a particular semantic mapping.

Optionally, the processor is configured to:

- generate the mapping function using a machine learning technique, and
- use the user feedback data as learning feedback in the machine learning technique.

Machine learning techniques are particularly well suited for generating the mapping function and adjusting the mapping function in view of the user feedback, as they avoid the need for devising complex heuristics to generate the mapping function.

For example, the machine learning technique may be a neural network, and the processor may be configured to adjust one or more weights of the neural network on the basis of the user feedback data.

Optionally, the source database and the target database are clinical databases.

It will be appreciated by those skilled in the art that two or more of the above-mentioned embodiments, implementations, and/or optional aspects of the invention may be combined in any way deemed useful.

Modifications and variations of the method and/or the computer readable media, which correspond to the described modifications and variations of the system, can be carried out by a person skilled in the art on the basis of the present description.

BRIEF DESCRIPTION OF THE DRAWINGS

These and other aspects of the invention will be apparent from and elucidated further with reference to the embodiments described by way of example in the following description and with reference to the accompanying drawings, in which

Fig. 1 shows a system for mapping data models;

Fig. 2 shows a database structured according to a data model, with the comprising data elements, data records and data items;

Fig. 3 shows a visualization of source data elements, target data elements, a mapping between said data elements, and confidence values associated with the mapping;

Fig. 4 shows another visualization of a mapping function and associated confidence values, with the user being prompted to provide user feedback;

Fig. 5 shows a visualization of data conversion as part of the mapping;

Fig. 6 shows an example of a neural network;

Fig. 7 illustrates the mapping function being generated using the neural network and using user feedback data as learning feedback; and

Fig. 8 shows a computer readable medium comprising instructions for causing a processor system to perform the method.

It should be noted that the figures are purely diagrammatic and not drawn to scale. In the figures, elements which correspond to elements already described may have the same reference numerals.

List of reference numbers

The following list of reference numbers is provided for facilitating the interpretation of the drawings and shall not be construed as limiting the claims.

020 source database
022 source data
040 target database
042 target data

060 display
062 display data

080 user input device
082 user feedback data

100 system for mapping data models
120 input interface
122 data communication
140 processor
142 data communication
160 user interface subsystem
162 display processor
164 user input interface

200, 202 source data elements
210 source data items
220, 222 target data elements
240, 242 mapping with confidence values

300-310 visualization of mapping function and confidence values
320 visualization of data conversion

400 neural network
410 input nodes layer
420 hidden nodes layer
430 output nodes layer

500 generate mapping using neural network
510 obtain user feedback indicative of desired mapping
520 mapping function
530 desired mapping function
540 feedback data to adapt weights in neural network

600 computer readable medium

610 non-transitory data representing instructions

DETAILED DESCRIPTION OF EMBODIMENTS

Fig. 1 shows a system 100 for mapping data models. The system 100 is shown to comprise an input interface 120 configured to access, via respective data communication 022, 042, a source database 020 structured according to a source data model, and a target database 040 structured according to a target data model. For example, the input interface 120 may enable the system 100 to read and/or write data items from both databases 020, 040. Although not explicitly shown in Fig. 1, the input interface 120 is further configured to access a source data model description describing the source data model, and a target data model description describing the target data model. For example, both descriptions may be stored on, and thus accessed from, the respective databases 020, 040. In general, the input interface 120 may take various forms, such as a network interface to a local or wide area network, such as the Internet, a storage interface to an internal or external data storage, etc.

The system 100 is further shown to comprise a processor 140 configured to internally communicate with the input interface 120 via data communication 122, and a user interface subsystem 160 which comprises a display processor 162 and a user input interface 164 which is configured to internally communicate with the processor 140 via data communication 142. The processor 140 may be configured to, during an operation of the system 100, generate a mapping function mapping the source data model to the target data model, for example on the basis of the respective data model descriptions and/or on the basis of the data items comprised in both databases, thereby enabling data items to be loaded from the source database into the target database on the basis of the mapping function. Here, the mapping function may comprise components for mapping respective data elements of the source data model to the target model. The processor 140 may be further configured to, during the operation of the system 100, determine confidence values of the mapping represented by individual components of the mapping function.

Moreover, the display processor 162 may be configured to, during the operation of the system 100, generate display data 062 comprising a visualization of the mapping function and the confidence values. The user input interface 164 may be configured to, during the operation of the system 100, obtain user feedback data 082 indicative of a correction of at least one component of the mapping function. As shown in Fig. 1, such display data 062 may be provided to a display 060, while the user feedback data 082 may be

received from a user input device 080 operable by the user, e.g., a computer mouse, touch screen, keyboard, etc. The processor 140 may be further configured to, during the operation of the system 100, adjust the mapping function on the basis of the user feedback data 082.

The mapping function is thereby generated by the system 100 in a 'hybrid' manner, which is less labor intensive and thereby less error prone as the user does not have to generate the mapping function from 'the ground up'. Nevertheless, a similar or higher quality as a manually generated mapping function may be obtained based on the user selectively correcting the (typically sparse) erroneous mappings of data elements.

The operation of the system 100, including various optional aspects thereof, will be further described with reference to Figs. 2-7.

In general, the system 100 may be embodied as, or in, a single device or apparatus, such as a workstation. The device or apparatus may comprise one or more (micro)processors which execute appropriate software. The software may have been downloaded and/or stored in a corresponding memory, e.g., a volatile memory such as RAM or a non-volatile memory such as Flash. Alternatively, the units of the system may be implemented in the device or apparatus in the form of programmable logic, e.g., as a Field-Programmable Gate Array (FPGA). In general, each unit of the system may be implemented in the form of a circuit. It is noted that the system 100 may also be implemented in a distributed manner, e.g., involving different devices or apparatuses. For example, the distribution of the system 100 may be in accordance with a client-server model.

Fig. 2 shows a database structured according to a data model, which illustrates the terms 'data element' and 'data item' as used in relation with the source database and the target database. The database may thus be an example of a source database 020, and may comprise a plurality of data elements 200, being represented by the labels '*PatientID*', '*Date of Birth*', '*pGleasonSum*' and '*PSA*' for the columns A-D. The rows, e.g., row 2 and 3, may each represent data records in the database, each comprising data items 210 for the respective data elements, e.g., the alphanumerical string '*John Doe*' as a value for the data element '*Patient ID*', the date 1-1-1960 for the data element '*Date of Birth*', the numeric value 7.00 for the data element '*pGleasonSum*' and the numeric value 4.00 for the data element '*PSA*'.

Fig. 3 shows an example of an output of the system 100 of Fig. 1, namely a visualization 300 of an automatically generated mapping function between two data models, and confidence values which are associated with the mapping function. In particular, the visualization 300, which may be shown to the user on a display, may comprise visual representations of the source data elements 200 of Fig. 2, visual representations of the target

data elements 220 and the mapping 240 between the data elements including confidence values. Effectively, the mapping 240 may be a 'proposed' or 'estimated mapping with a confidence of the proposal or estimate being indicated to the user. For example, it is shown in Fig. 3 that the system has mapped the source data element '*PatientID*' to the target data element '*PatientID*' with a confidence of 100%. It is further shown that the system has mapped the source data element '*Date of Birth*' to the target data element '*DOB*' with a confidence of 95%, and that it has mapped the source data element '*pGleasonSum*' to the target data element '*Pathological Gleason*' with a confidence of 60%. Lastly, it can been seen from the visualization 300 that the system has not mapped the source data element '*PSA*', nor generated a mapping to the target data element '*Prostate Volume*'.

Fig. 4 shows another visualization 310 of an automatically generated mapping function and associated confidence values. In this example, the mapping function is represented by horizontal lines 242 comprising the confidence value (as a percentage), with the source data elements 202 and the target data elements 222 being ordered by row such that data elements which are mapped, e.g., by way of a horizontal line, are comprised in a same row. Fig. 4 further shows the user being prompted to provide user feedback, e.g., by the prompt "*What do you want to do? (a)ccept row, (r)eject row, (n)ew match, (e)nd mapping, (all) accept*". In this example, the user is thus enabled to accept a particular mapping by accepting the therewith associated row, reject a particular mapping by rejecting the associated row, specifying a new match, accept all mappings, and end the mapping generation. As will be also further explained with reference to Fig. 6, the user feedback may enable the system to improve the mapping function.

A specific example may be that if data elements have been mapped with a high confidence level, but their mappings are subsequently being overridden by the user, their mapping may be 'blacklisted'. In this case, the mapping may not only be rejected for this occurrence, but this may also result in a lower confidence level in future occurrences. On the other hand, if data elements have been mapped with a low confidence level, but their mappings have been approved by the user, their mapping might be 'whitelisted'. In this case, the mapping is not only accepted for this occurrence, but it also results in a higher confidence level in future occurrences.

In general, the mapping function may be generated using data-driven mapping and semantic mapping. Semantic mapping may make use of a synonym list which may be consulted to look-up data element synonyms. For example, if the source database comprises '*DOB*' as data element whereas the target database comprises '*DateOfBirth*' as data element,

both data elements may be mapped by the sematic mapping if they are listed as synonyms in the synonym list. In general, semantic mapping may make use of an ontology or data standard, such as CDISC. A limitation of semantic mapping is that it may only determine more-or-less 'exact' matches between data elements, e.g., as represented by columns of data, and will not discover any transformation logic or exceptions between the data elements.

Here, data-driven mapping may play a role. Date-driven mapping may involve simultaneously evaluating actual data items in two databases using heuristics and statistics to automatically discover complex mappings between the two databases. Examples of such data-driven mapping include the following (here, the examples from the prostate cancer field). As a first non-limiting example, a value of a data item may indicate what data type it is, especially when there only exist a limited number of predetermined values. For example, the data item values '*cT2a*' and '*cT3b*' may indicate that the data element pertains to cT values. The range of numerical values may also help to identify the data element. For example, a range from 2 to 10 may indicate that the data element pertains to Gleason. Similarly, the distribution of numerical values may help to identify the data element. For example, in case of the data element pertaining to Gleason, the numerical values would typically have a normal distribution around 6 or 7. Also data items associated with other data elements may be used to identify the data element. For example, in case a '*Gleason1*' column contains 4 as data item value and a '*Gleason2*' column contains 3 as data item value, a column containing 7 as data item value may be identified as a GleasonSum. The system of Fig. 1 may use data-driven mapping as an alternative, or in addition to semantic mapping, for example, when the semantic mapping does not meet a mapping criterion, such as a confidence value threshold, or to verify the correctness of the semantic mapping.

Fig. 5 shows a visualization 320 of data conversion. Such data conversion may be part of the automatic generation of the mapping function, in that it may be automatically determined that a source data element may be mapped to a particular target element, but that the data items associated with the source data element may be represented differently than the data items associated with the target data element. For example, it may be determined that the source data items are expressed in a different unit than the target data items, that the notation of time and/or date may be different, etc. As such, although the data items may principally be of a same type, there may be a conversion required. Fig. 5 shows two type of such conversions having been automatically detected by the system, e.g., using techniques known per se in the art of data conversion, but of which the correctness is verified with the user.

In particular, it is shown that the source data element '*Date_birth*' has been mapped to the target data element '*Year of Birth*', but that it has been recognized by the system that the source data element represents a date whereas the target data element represents a year. Accordingly, a conversion is proposed from DD-MM-YYYY to YYYY, which is verified with the user with the prompt "*Are you sure you accept this conversion? (y)es to accept, (n)o to keep the data unchanged*". It is further shown that the source data element '*Length*' has been mapped to the target data element '*Height at diagnosis*', but that it has been recognized by the system that the source data element represents a decimal height in meters whereas the target data element represents an integer height in centimeters. Accordingly, a conversion is proposed in which the decimal height is multiplied by 100.0, which is verified with the user with the prompt "*All data will be multiplied with: 100.0 and will have no decimals. Do you want to accept this conversion? (y)es to accept, (n)o*".

It is noted that the user interface subsystem may be configured to enable the user interaction as described with reference to Figs. 3-5.

In general, the processor of the system of Fig. 1 may be configured to generate the mapping function using a machine learning technique, and to use the user feedback data as learning feedback in the machine learning technique. Any suitable machine learning technique may be used, including but not limited to a neural network, which is shown in its basic form in Fig. 6. Here, the neural network 400 is shown to comprise an input nodes layer 410, a hidden nodes layer 420 connected to the input nodes layer 410 via weights w_1^j, and an output nodes layer 430 connected to the hidden nodes layer 420 via weights w_2^j. An output of the neural network may be obtained in the form of $y_i = f(w_i^1 x_1 + w_i^2 x_2 + \ldots + w_i^m x_m) = f(\sum_{j=1}^{m} w_i^j x_j)$. The input layer may consist of one or more of: semantic features (e.g. synonyms, ontology features), data-driven features (data type, range of numerical values, data element distribution) and whitelist/blacklist rules, all belonging to a data element of a source database that needs to be mapped to a data element of a target database. The output layer may consist of the confidence levels of all the possible mappings (of the to-be-mapped data element) to the data elements of the target database and the weights may represent how much should the input features contribute to these confidence levels. As such, as also illustrated in Fig. 7, the neural network may generate 500 a mapping function 520 which may be visualized to a user. The user may then provide 510 user feedback indicative of desired mapping function 530. A difference between the generated mapping function 520 and the desired mapping function 530 may be determined, which may then represent feedback data 540 which may be used to adapt the weights in the neural network.

In general, a method of mapping data models may comprise, in an operation titled "ACCESSING DATABASES", accessing a source database structured according to a source data model, and a target database structured according to a target data model. The method may further comprise, in an operation titled "GENERATING MAPPING FUNCTION", generating a mapping function mapping the source data model to the target data model to enable loading data items from the source database into the target database on the basis of the mapping function, wherein the mapping function comprises components for mapping respective data elements of the source data model to the target model. The method may further comprise, in an operation titled "DETERMINING CONFIDENCE VALUES", determining confidence values of the mapping represented by respective components of the mapping function. The method may further comprise, in an operation titled "GENERATING VISUALIZATION OF MAPPING FUNCTION AND CONFIDENCE VALUES", generating display data comprising a visualization of the mapping function and the confidence values. The method may further comprise, in an operation titled "RECEIVING CORRECTION OF MAPPING FUNCTION", obtaining user feedback data indicative of a correction of at least one component of the mapping function. The method may further comprise, in an operation titled "ADJUSTING MAPPING FUNCTION", adjusting the mapping function on the basis of the user feedback data.

The method may be implemented on a computer as a computer implemented method, as dedicated hardware, or as a combination of both. As also illustrated in Fig. 8, instructions for the computer, e.g., executable code, may be stored on a computer readable medium 600, e.g., in the form of a series 610 of machine readable physical marks and/or as a series of elements having different electrical, e.g., magnetic, or optical properties or values. The executable code may be stored in a transitory or non-transitory manner. Examples of computer readable mediums include memory devices, optical storage devices, integrated circuits, servers, online software, etc. Fig. 8 shows an optical disc 600.

It will be appreciated that, in accordance with the abstract of the present specification, a system and method are provided for generating a mapping function for use in loading data from a source database, which is structured in accordance with a source data model, into a target database, which is structured in accordance with a target data model. The mapping function is automatically generated, e.g., on the basis of descriptions of both data models and/or the data items comprised in the both databases, with the mapping comprising components for mapping respective data elements of the source data model to the target model. To improve the quality of the mapping function, confidence values of the mapping are

determined and visualized together with the mapping. User feedback is obtained indicative of a correction of at least one component of the mapping function. In response, the mapping function is adjusted on the basis of the user feedback data. Advantageously, the mapping function may be generated in a 'hybrid' manner, which is less labor intensive and less error prone than having to generate the mapping function by hand.

Examples, embodiments or optional features, whether indicated as non-limiting or not, are not to be understood as limiting the invention as claimed.

It will be appreciated that the invention also applies to computer programs, particularly computer programs on or in a carrier, adapted to put the invention into practice. The program may be in the form of a source code, an object code, a code intermediate source and an object code such as in a partially compiled form, or in any other form suitable for use in the implementation of the method according to the invention. It will also be appreciated that such a program may have many different architectural designs. For example, a program code implementing the functionality of the method or system according to the invention may be sub-divided into one or more sub-routines. Many different ways of distributing the functionality among these sub-routines will be apparent to the skilled person. The sub-routines may be stored together in one executable file to form a self-contained program. Such an executable file may comprise computer-executable instructions, for example, processor instructions and/or interpreter instructions (e.g. Java interpreter instructions). Alternatively, one or more or all of the sub-routines may be stored in at least one external library file and linked with a main program either statically or dynamically, e.g. at run-time. The main program contains at least one call to at least one of the sub-routines. The sub-routines may also comprise function calls to each other. An embodiment relating to a computer program product comprises computer-executable instructions corresponding to each processing stage of at least one of the methods set forth herein. These instructions may be sub-divided into sub-routines and/or stored in one or more files that may be linked statically or dynamically. Another embodiment relating to a computer program product comprises computer-executable instructions corresponding to each means of at least one of the systems and/or products set forth herein. These instructions may be sub-divided into sub-routines and/or stored in one or more files that may be linked statically or dynamically.

The carrier of a computer program may be any entity or device capable of carrying the program. For example, the carrier may include a data storage, such as a ROM, for example, a CD ROM or a semiconductor ROM, or a magnetic recording medium, for example, a hard disk. Furthermore, the carrier may be a transmissible carrier such as an

electric or optical signal, which may be conveyed via electric or optical cable or by radio or other means. When the program is embodied in such a signal, the carrier may be constituted by such a cable or other device or means. Alternatively, the carrier may be an integrated circuit in which the program is embedded, the integrated circuit being adapted to perform, or used in the performance of, the relevant method.

It should be noted that the above-mentioned embodiments illustrate rather than limit the invention, and that those skilled in the art will be able to design many alternative embodiments without departing from the scope of the appended claims. In the claims, any reference signs placed between parentheses shall not be construed as limiting the claim. Use of the verb "comprise" and its conjugations does not exclude the presence of elements or stages other than those stated in a claim. The article "a" or "an" preceding an element does not exclude the presence of a plurality of such elements. The invention may be implemented by means of hardware comprising several distinct elements, and by means of a suitably programmed computer. In the device claim enumerating several means, several of these means may be embodied by one and the same item of hardware. The mere fact that certain measures are recited in mutually different dependent claims does not indicate that a combination of these measures cannot be used to advantage.

CLAIMS:

1. A system (100) for mapping data models, comprising:

- an input interface (120) configured to access:
  - a source database (020) structured according to a source data model,
  - a target database (040) structured according to a target data model, and
- a processor (140) configured to generate a mapping function (240, 242) mapping the source data model to the target data model to enable loading data items (210) from the source database into the target database on the basis of the mapping function, wherein the mapping function comprises components for mapping respective data elements (200, 220) of the source data model to the target model,

wherein the processor is configured for determining confidence values of the mapping represented by respective components of the mapping function as a result of the mapping function being automatically generated; and

- a user interface subsystem (160) comprising:
  - a display processor (162) configured to generate display data (062) comprising a visualization (300, 310) of the mapping function and the confidence values, and
  - a user input interface (164) configured to obtain user feedback data (082) indicative of a correction of at least one component of the mapping function;

and wherein:

- the processor is configured to adjust the mapping function on the basis of the user feedback data,

wherein the processor (140) is configured to generate the mapping function (240, 242) using both data-driven mapping and semantic mapping.

2. The system (100) according to claim 1, wherein the processor (140) is configured to generate the mapping function (240, 242) on the basis of a source data model description describing the source data model and a target data model description describing the target data model.

3. The system (100) according to claim 2, wherein the data-driven mapping is based on at least one of:

- a value of a data item (210) of a source data element (200, 202),
- a relationship between the value of the data item of the source data element and data items of one or more other source data elements,
- a range of values of data items of the source data element, and
- a distribution of values of data items of the source data element,

being indicative of a target data element (220, 222).

4. The system (100) according to any one of claims 1 or 3, wherein the processor (140) is configured to, when generating the mapping function (240, 242), use data-driven mapping for mapping a source data element (200, 202) to a target data element (220, 222) when the semantic mapping does not meet a mapping criterion.

5. The system (100) according to any one of claims 1 to 4, wherein the processor (140) is configured to:

- generate the mapping function (240, 242) using a machine learning technique (400), and
- use the user feedback data (082) as learning feedback in the machine learning technique.

6. The system (100) according to claim 5, wherein the machine learning technique is a neural network (400), and wherein the processor is configured to adjust one or more weights of the neural network on the basis of the user feedback data (082).

7. The system (100) according to any one of the above claims, wherein the source database (020) and the target database (040 are clinical databases.

8. Use of the system according to any one of claims 1 to 7 to extract, transform and load of the data items from the source database (040) into the target database (040).

9. A method of mapping data models, comprising:

- accessing:

- a source database structured according to a source data model,
- a target database structured according to a target data model, and
- generating a mapping function (500) using both data-driven mapping and semantic mapping for mapping the source data model to the target data model to enable loading data items from the source database into the target database on the basis of the mapping function, wherein the mapping function comprises components for mapping respective data elements of the source data model to the target model,
- determining confidence values of the mapping represented by respective components of the mapping function as a result of the mapping function being automatically generated;
- generating display data comprising a visualization of the mapping function and the confidence values;
- obtaining user feedback data (510) indicative of a correction of at least one component of the mapping function; and
- adjusting the mapping function on the basis of the user feedback data.

10. The method according to claim 9, further comprising:
- loading the data items from the source database into the target database on the basis of the mapping function.

11. A computer readable medium (600) comprising transitory or non-transitory data (610) representing instructions arranged to cause a processor system to perform the method according to any one of claims 9 or 10.

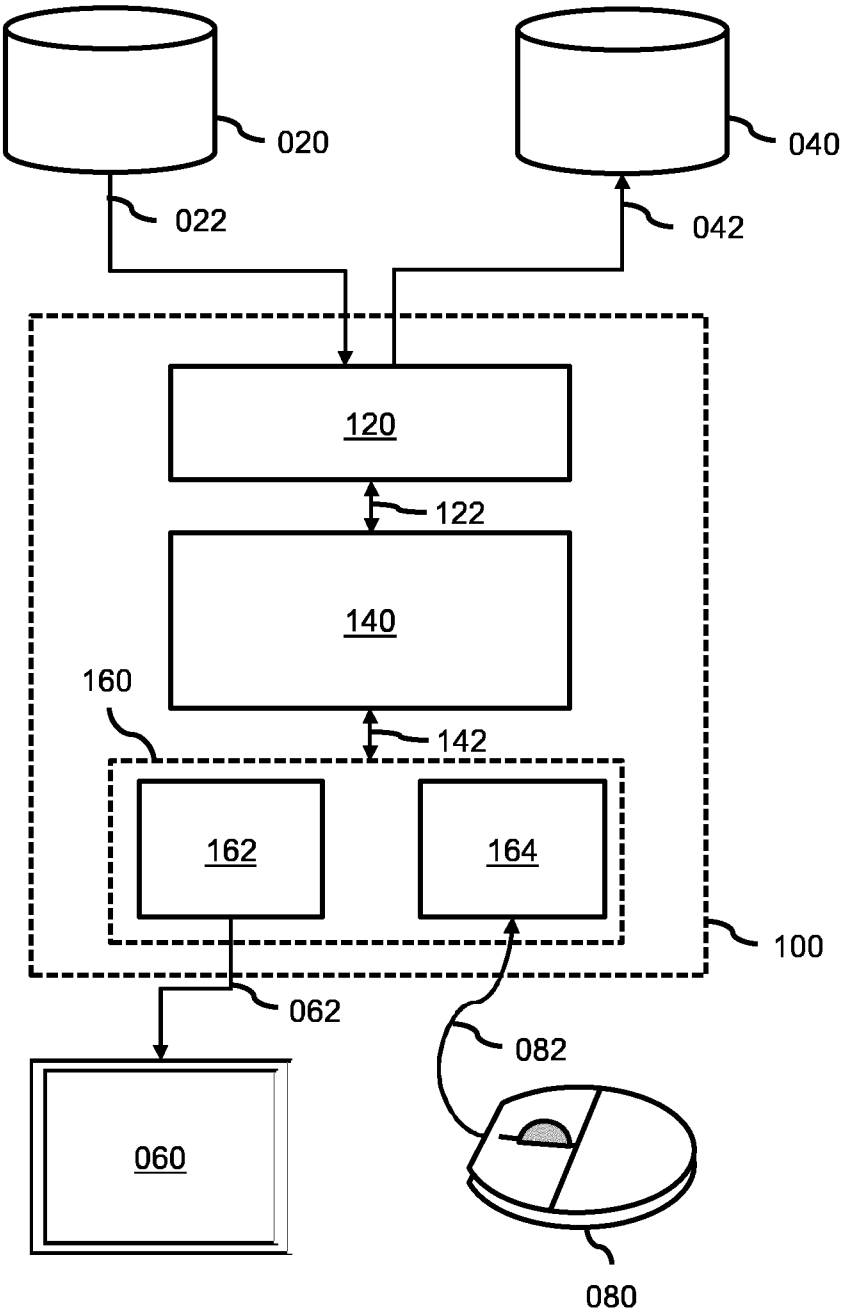

Fig. 1

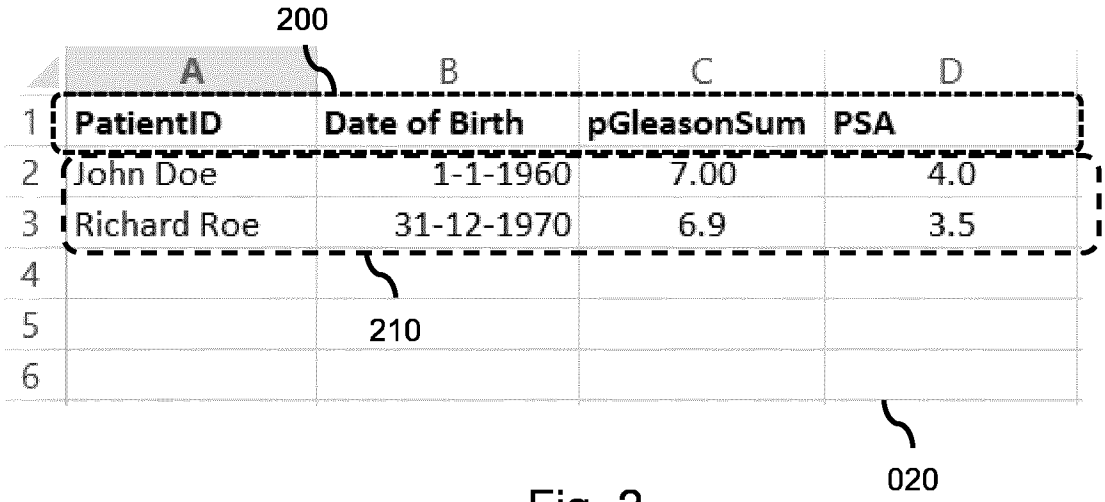
200
A
B
C
D
1
PatientID
Date of Birth
pGleasonSum
PSA
2
John Doe
1-1-1960
7.00
4.0
3
Richard Roe
31-12-1970
6.9
3.5
4
5
6
210
020

Fig. 2

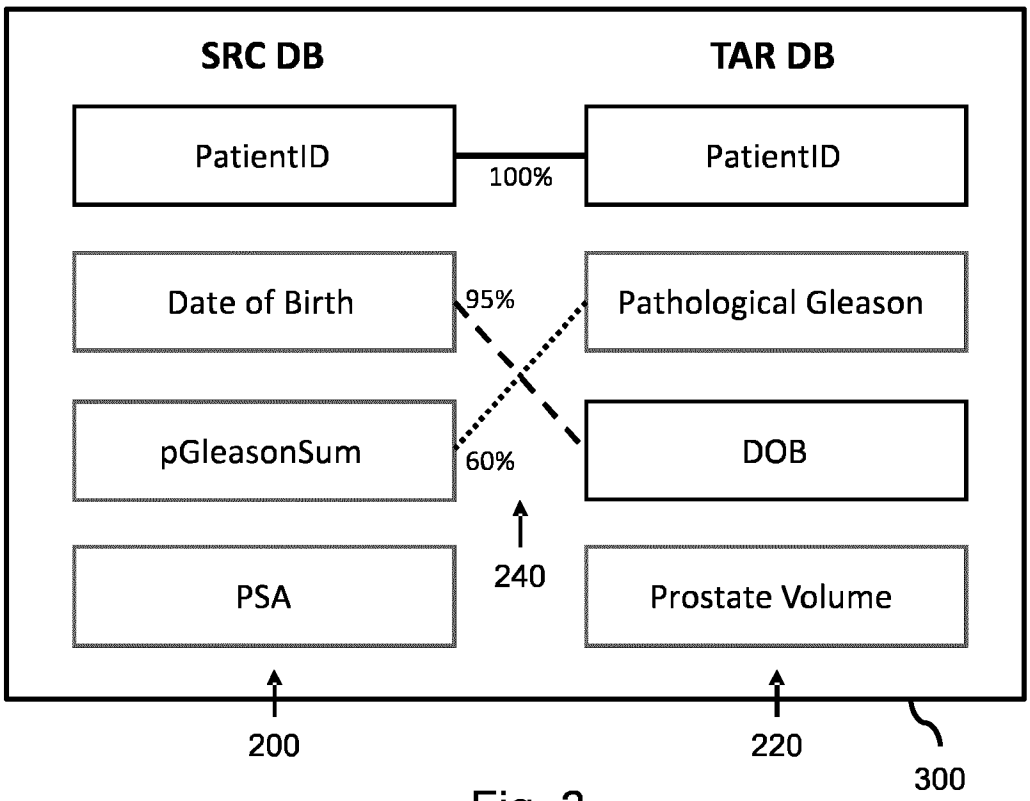
SRC DB
TAR DB
PatientID
100%
PatientID
Date of Birth
95%
Pathological Gleason
pGleasonSum
60%
DOB
240
PSA
Prostate Volume
200
220
300

Fig. 3

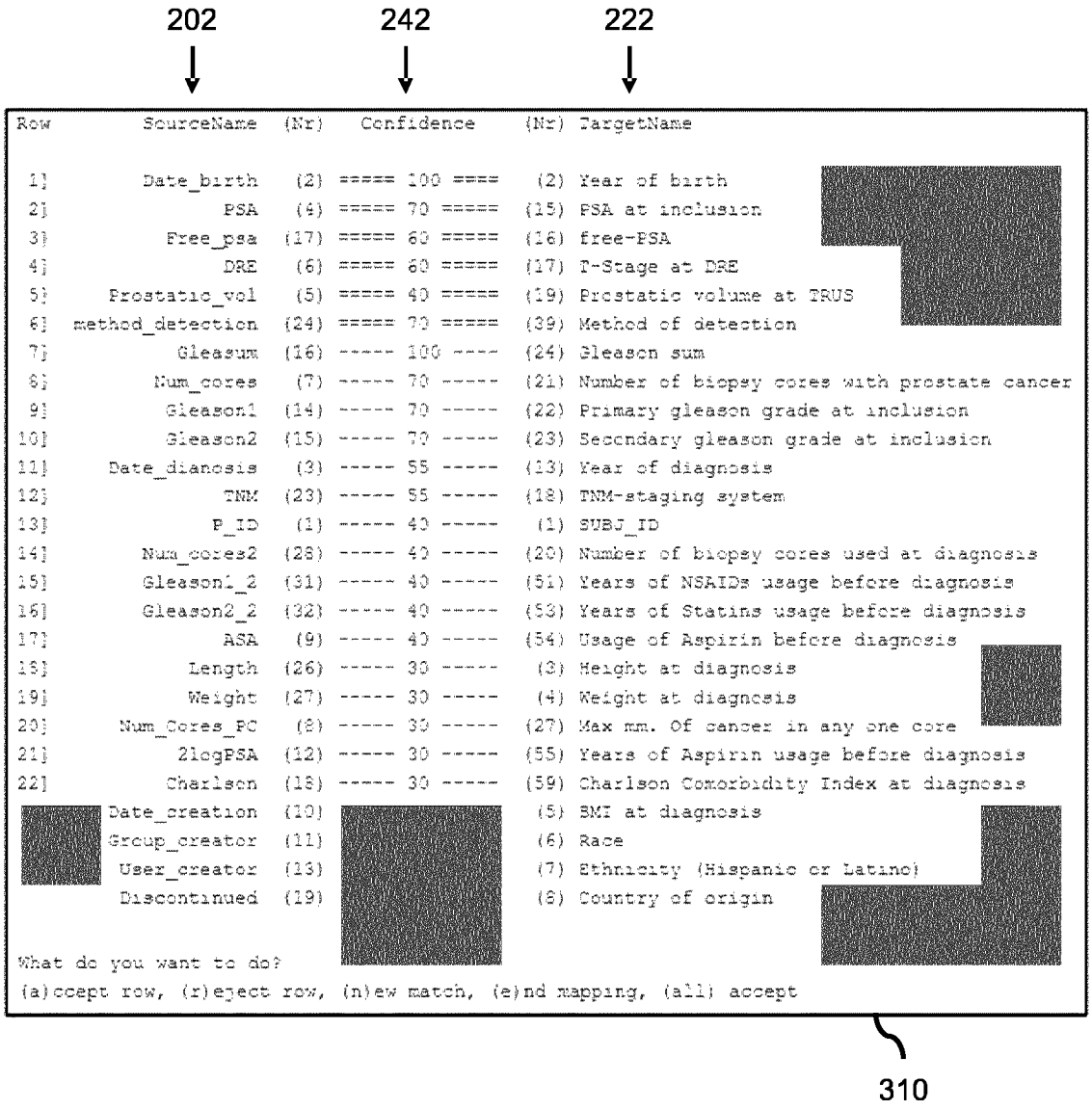
202
242
222
Row SourceName (Nr) Confidence (Nr) TargetName
1] Date_birth (2) ===== 100 ==== (2) Year of birth
2] PSA (4) ===== 70 ===== (15) PSA at inclusion
3] Free_psa (17) ===== 60 ===== (16) free-PSA
4] DRE (6) ===== 60 ===== (17) T-Stage at DRE
5] Prostatic_vol (5) ===== 40 ===== (19) Prostatic volume at TRUS
6] method_detection (24) ===== 70 ===== (39) Method of detection
7] Gleasum (16) ----- 100 ---- (24) Gleason sum
8] Num_cores (7) ----- 70 ----- (21) Number of biopsy cores with prostate cancer
9] Gleason1 (14) ----- 70 ----- (22) Primary gleason grade at inclusion
10] Gleason2 (15) ----- 70 ----- (23) Secondary gleason grade at inclusion
11] Date_diagnosis (3) ----- 55 ----- (13) Year of diagnosis
12] TNM (23) ----- 55 ----- (18) TNM-staging system
13] P_ID (1) ----- 40 ----- (1) SUBJ_ID
14] Num_cores2 (28) ----- 40 ----- (20) Number of biopsy cores used at diagnosis
15] Gleason1_2 (31) ----- 40 ----- (51) Years of NSAIDs usage before diagnosis
16] Gleason2_2 (32) ----- 40 ----- (53) Years of Statins usage before diagnosis
17] ASA (9) ----- 40 ----- (54) Usage of Aspirin before diagnosis
18] Length (26) ----- 30 ----- (3) Height at diagnosis
19] Weight (27) ----- 30 ----- (4) Weight at diagnosis
20] Num_Cores_PC (8) ----- 30 ----- (27) Max mm. Of cancer in any one core
21] 2logPSA (12) ----- 30 ----- (55) Years of Aspirin usage before diagnosis
22] Charlson (18) ----- 30 ----- (59) Charlson Comorbidity Index at diagnosis
Date_creation (10) (5) BMI at diagnosis
Group_creator (11) (6) Race
User_creator (13) (7) Ethnicity (Hispanic or Latino)
Discontinued (19) (8) Country of origin
What do you want to do?
(a)ccept row, (r)eject row, (n)ew match, (e)nd mapping, (all) accept
310

Fig. 4

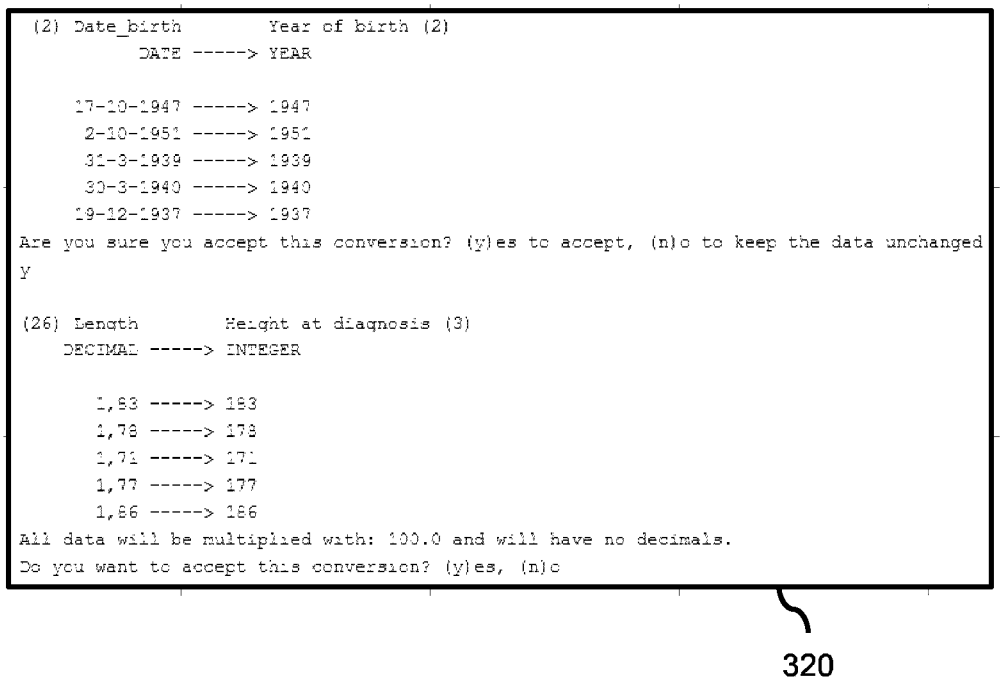
(2) Date_birth Year of birth (2)
DATE -----> YEAR
17-10-1947 -----> 1947
2-10-1951 -----> 1951
31-3-1939 -----> 1939
30-3-1940 -----> 1940
19-12-1937 -----> 1937
Are you sure you accept this conversion? (y)es to accept, (n)o to keep the data unchanged
y
(26) Length Height at diagnosis (3)
DECIMAL -----> INTEGER
1,83 -----> 183
1,78 -----> 178
1,71 -----> 171
1,77 -----> 177
1,86 -----> 186
All data will be multiplied with: 100.0 and will have no decimals.
Do you want to accept this conversion? (y)es, (n)o
320

Fig. 5

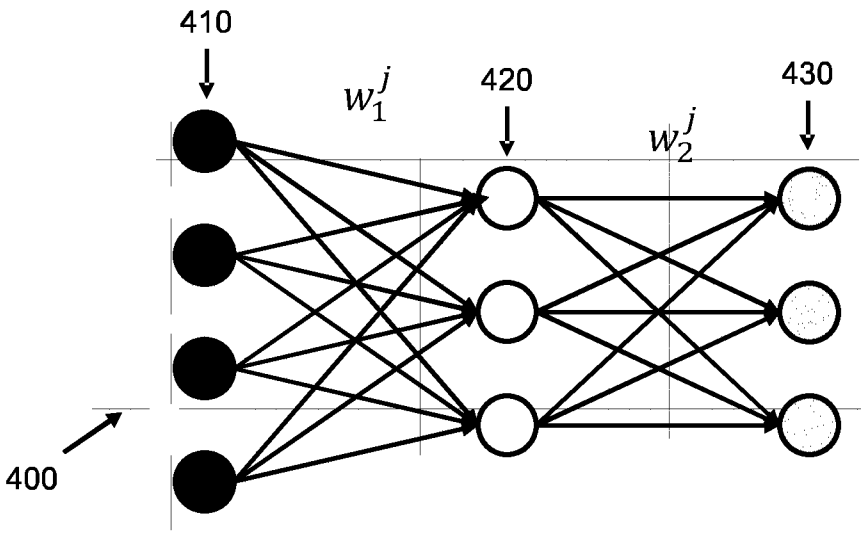
410
w_1^j
420
w_2^j
430
400

Fig. 6

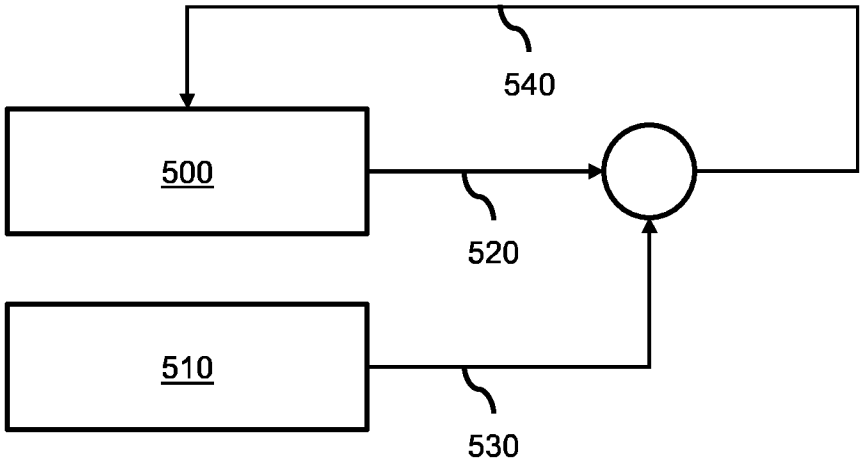

Fig. 7

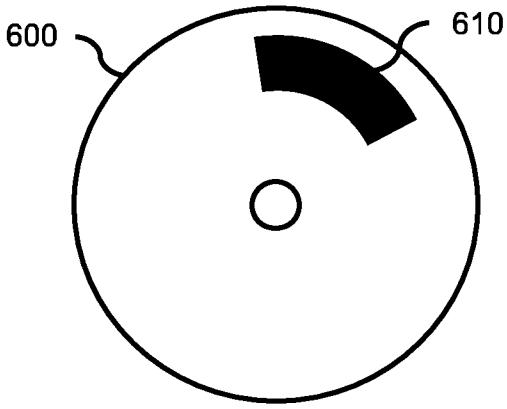

Fig. 8

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2017/056898

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F19/00
ADD. G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2008/021912 A1 (SELIGMAN LEONARD J [US] ET AL) 24 January 2008 (2008-01-24) abstract; claims 1,5,16; figures 1-6 paragraph [0004] - paragraph [0005] paragraph [0013] - paragraph [0017] paragraph [0032] - paragraph [0034] paragraph [0037] - paragraph [0040] paragraph [0064] - paragraph [0065] paragraph [0071] paragraph [0079] - paragraph [0083] ----- -/--	1-11

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
8 June 2017	22/06/2017

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Filloy García, E

1

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2017/056898

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2016/063209 A1 (MALAVIYA SANJAY [CA]) 3 March 2016 (2016-03-03) abstract; claims 1-4 paragraph [0041] paragraph [0046] paragraph [0059] paragraph [0063] - paragraph [0070] paragraph [0173] - paragraph [0177] -----	1-11
X	US 2011/295866 A1 (FOT DMITRIY [KZ] ET AL) 1 December 2011 (2011-12-01) abstract; figures 1-5 paragraph [0017] - paragraph [0018] paragraph [0031] - paragraph [0034] paragraph [0037] paragraph [0041] - paragraph [0043] -----	1-11
X	US 2014/095205 A1 (FAROOQ FAISAL [US] ET AL) 3 April 2014 (2014-04-03) abstract; claim 1; figures 1,2 paragraph [0006] - paragraph [0009] paragraph [0023] - paragraph [0028] paragraph [0046] - paragraph [0048] paragraph [0052] - paragraph [0057] paragraph [0061] - paragraph [0066] -----	1-11

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2017/056898

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2008021912 A1	24-01-2008	NONE	
US 2016063209 A1	03-03-2016	CA 2902105 A1	28-02-2016
		US 2016063209 A1	03-03-2016
US 2011295866 A1	01-12-2011	US 2011295866 A1	01-12-2011
		US 2012254205 A1	04-10-2012
US 2014095205 A1	03-04-2014	NONE	